Proceedings of CONF-MLA 2025 Symposium: Intelligent Systems and Automation: AI Models, IoT, and Robotic Algorithms
DOI: 10.54254/2755-2721/2025.LD28453

# *Machine Learning-Based Forecasting of Renewable Power Generation Using Meteorological Data*

**Lyu Han**

*Department of Statistics, University of Illinois Urbana-Champaign, Illinois, America*
*hanlyu2@illinois.edu*

***Abstract.*** This study tackles short-term wind power forecasting using hourly data (2017–2021) from four utility-scale sites. An end-to-end machine learning pipeline is constructed with strict quality control and physics-informed features, including u/v wind vector decomposition, nonlinear wind-speed terms, and compact calendar encodings (hour and month). Model evaluation combines rolling-origin time splits with leave-one-site-out (LOSO) cross-site testing. Stable feature importance analysis yields a Top-27 feature set, ensuring comparability across sites and deployment readiness. In pooled training, LightGBM performs best (RMSE ≈ 0.161, $R^2$ ≈ 0.45), while results reveal strong site heterogeneity. LOSO testing improves generalization at lower-skill sites, and adding time encodings further tightens per-site fits ($R^2$ ≈ 0.98). The contribution is a reproducible forecasting workflow that balances physical interpretability with predictive accuracy, a lean, transferable feature set, and a rigorous evaluation protocol that separates temporal from spatial generalization. Findings inform operational forecasting for wind assets and offer a practical blueprint for scaling predictive maintenance and dispatch decisions across diverse wind regimes.

***Keywords:*** Renewable energy forecasting, Wind power prediction, Machine learning models, Cross-site generalization.

## 1. Introduction

With wind generation continuing to grow, system operators are looking more and more to accurate forecasts from minutes to days into the future to help them schedule flexible resources, size reserves and more, to maintain reliability under weather-dependent conditions [1]. On the meteorological side, continuous development of numerical weather prediction (NWP) at convection-allowing scales and focused boundary-layer improvements have measurably enhanced hub-height wind representation and event-specific bias correction, creating a stronger foundation for energy forecasting when models are properly post-processed [2, 3]. For historical training and climatological context, global reanalyses such as ERA5 provide consistent multidecadal datasets [4], while energy-oriented resources like NREL's WIND Toolkit pair meteorological variables with power time series across many U.S. sites to enable reproducible studies [5].

Methodologically, the field has evolved from purely physical and classical statistical approaches toward data-driven machine learning (ML) and deep learning (DL), which better capture nonlinear

DOI: 10.54254/2755-2721/2025.LD28453

relationships and multi-variable interactions linking atmospheric drivers to turbine output [6, 7]. Gradient-boosted decision trees such as XGBoost and LightGBM have repeatedly delivered strong accuracy and computational efficiency on multivariate regression problems relevant to renewable forecasting [8, 9]. These gains, however, hinge on rigorous data preparation—such as time alignment, gap handling, and outlier screening—and on evaluation protocols that respect temporal and spatial dependence [10-12].

Because the wind vector and air density tightly constrain turbine power, meteorology-aware features are essential. Representing direction as a circular quantity (e.g., via angles or sine–cosine embeddings) avoids discontinuities at 0°/360° and supports stable learning; equivalently, decomposing winds into u–v components often improve performance [13, 14]. Effects from thermodynamics are captured with density terms from temperature, humidity, and pressure [15], and simple nonlinear expansions of wind speed crudely approximate the curvature of turbine power curves in the partial-load regime [7]. Short lags of power and key meteorological variables also leverage the importance of local persistence for short-term horizons [16].

Evaluation design is just as crucial. Random k-fold cross-validation is unrealistically optimistic when observations are temporally or spatially autocorrelated. Best practice instead involves rolling-origin evaluation for time series and explicit spatial holdouts for generalizing across locations; these principles have gained widespread acceptance in forecasting and geospatial ML [16-18]. These protocols are compatible with energy-forecasting benchmarks that emphasize transparent, comparable metrics and a careful separation of training and testing periods [19].

People operationalize those principles at four utility-scale wind sites (2017–2021) with hourly power and co-located meteorological data. The paper first standardizes and quality-controls the data —including UTC alignment, gap handling, anomaly screening, and directional encodings— following open operational-analysis practices [10, 12]. Next, we employ linear and regularized baselines, Random Forest, XGBoost, and LightGBM under two complementary validation schemes —rolling origin and leave-one-site-out (LOSO) —to capture both temporal and spatial generalization [17,18]. The analysis then compares pooled (multi-site) and site-specific training to examine conditions unger which aggregation is beneficial or detrimental. In addition, physics-guided feature design is implemented, incorporating wind-speed nonlinearities, vector wind components, thermo-humidity interactions as density proxies, and short lags, followed by importance-driven pruning to a compact Top-27 set [8, 9]. For higher-skill sites, seasonal expert models are leveraged to utilize regime structure and explicit time features (hour and month) to quantify their incremental value at the global level [16].

The contributions of this paper can be summarized as follows. First, we provide a transparent, end-to-end pipeline that seamlessly couples principled data preparation with time- and space-aware validation; Second, we derive an empirical map of where pooled models are preferable to site-specific models, alongside ablation analyses of seasonal specialization and physics-guided features; Third, we introduce a small and interpretable set of features that retain or improve upon the accuracy of full models while facilitating deployment. Furthermore, the design and evaluation metrics are directly informed by community guidance on rigorous evaluation and meteorology-aware modeling. Our proposed framework naturally extends to calibrated probabilistic forecasts in future work [19].

## 2. Literature review

Wind power forecasting (WPF) is commonly organized into three methodological families—physics/NWP pipelines, statistical time-series models, and machine learning (ML)/deep learning approaches [6]. On the meteorology side, operational NWP systems such as RAP/HRRR and the

DOI: 10.54254/2755-2721/2025.LD28453

WRF-ARW modeling system provide the backbone for hub-height wind guidance from intra-day to day-ahead horizons [2, 3], while ERA5 supplies multidecadal, physically consistent training data and climatology [4]. Energy-oriented datasets, such as NREL's WIND Toolkit, pair meteorological variables with plant-level power series, enabling reproducible evaluation across multiple sites [5]. For evaluation, best practice avoids random k-fold cross-validation due to temporal and spatial dependence; instead, rolling-origin testing and location-aware holdouts are recommended. In this paper, these recommendations are operationalized through Time Rolling and LOSO validation [16-18].

In the context of data-driven WPF, gradient-boosted trees such as XGBoost and LightGBM consistently report high accuracy on multivariate, nonlinear relationships between weather and power (i.e., in tasks like this paper). Such models have complemented (and, in some cases, outperformed) linear baselines and Random Forests [8, 9]. Robust preprocessing is critical to success, and open operational analysis practices, anomaly detection, and power curve cleaning are all standard methods to improve data fidelity before modeling begins [10-12]. Physics-informed feature design is a common practice: treat wind direction data as circular (via angle or sin/cos) or convert it to speed-direction coordinates to u/v components; add density-related features/terms informed by temperature, humidity, and pressure; and add simple wind-speed nonlinearities and short lags to capture site persistence [7, 13, 15]. Time features, such as hour or month, help to explicitly encode diurnal and seasonal cycles, while seasonal experts might also exploit structure in regime predictability [16, 20]. Finally, model interpretation and compactness are improved via tree-based gain importances and SHAP, which underpin the Top 27 feature shortlist presented in this paper while maintaining accuracy [8, 9, 21]. Reporting in this paper follows the norms of the energy-forecasting community to ensure transparent and comparable metrics [19].

## 3. Method

### 3.1. Data

This study analyzes four utility-scale wind sites (Loc1–Loc4) with hourly observations from 2017–2021. Each record contains a normalized turbine power output in [0,1][0,1] and a consistent set of co-located meteorological variables: near-surface temperature at 2 m (°F), relative humidity at 2 m (%), dew point at 2 m (°F), wind speed at 10 m and 100 m (m $s^{-1}$), wind direction at 10 m and 100 m (degrees, 0–360), and wind gusts reported as windgusts_10m (m $s^{-1}$). All four sites share the same schema: Time, temperature_2m, relativehumidity_2m, dewpoint_2m, windspeed_10m, windspeed_100m, winddirection_10m, winddirection_100m, windgusts_10m, Power.

In this study, Power is considered the target to be predicted, with all meteorological variables as candidate predictors. Physics-aware encodings are derived, when possible (e.g. radians and vector components from wind direction and wind speed), and time-series constructs are deferred to the modelling stage (e.g. lags), to prevent information leakage.

A uniform site-by-site data-quality pipeline is applied. The process begins with the standardization of timestamps to UTC and sort by site_id, Time. For each site, a complete hourly index covering 2017–2021 is constructed, and raw records are left-jioned to this index in order to expose gaps and duplicates. Exact duplicates are removed, while non-identical rows sharing the same timestamp are resolved by retaining the latest ingest record and flag the conflicts in the quality log. After alignment, conservative range and unit checks are conducted: Power $\in$ [0,1]; wind speeds $\geq$0; wind directions in [0,360]; relative humidity in [0, 100] %; and temperatures/dew points within broad, site-specific percentile envelopes (e.g., 0.1–99.9th) to guard against unit slips and sensor

DOI: 10.54254/2755-2721/2025.LD28453

spikes. Missingness is then categorized by gap length. Short gaps (≤3 consecutive hours) are filled by linear interpolation of meteorological variables along time within the same site, whereas Power is never interpolated and long outages remain missing for auditing and are excluded from model training.

To further eliminate records that pass simple bounds but are still implausible, a robust power-curve filter is applied. Within each site, a LOESS fit of Power ~ windspeed_100m is constructed using non-gap data, and observations whose absolute residual exceeds 3×MAD (median absolute deviation) are removed. This step targets hidden curtailment, icing, and sensor faults without over-trimming extremes. Because wind direction is circular, it is encoded in radians (winddirection_*_rad), and when beneficial, sinusoidal embeddings (sin, cos) or vector wind components ($u=v\cos\theta$, $v=v\sin\theta$ at 10 m and 100 m) are added to avoid the 0°/360° discontinuity. All cleaning operations are performed within each site to preserve spatial independence. Finally, a comprehensive data-quality report is generated for every site, summarizing missingness by column, the distribution of gap lengths, counts removed at each quality-control stage (duplicates, range violations, robust filter), and retained sample sizes with pre- and post-cleaning statistics. The cleaned, row-level dataset is saved as df_all_clean.csv, while the hourly feature table with engineered circular/vector encodings is provided as features_h1.csv.

## 3.2. Model families and unified comparison (pooled four sites)

### 3.2.1. Models and setup

This study benchmarks six learners on the pooled dataset that combines Loc1–Loc4 using the same feature set and the normalized target Power. The models include Linear regression, Ridge, Lasso, Random Forest, XGBoost, and LightGBM. Linear, Ridge, and Lasso use standardization (mean 0, variance 1) fit on the training fold only and applied to the test fold. Random Forest uses 500 trees, sqrt(p) features per split, and min_samples_leaf = 5. XGBoost uses learning rate 0.03, max_depth 6, subsample 0.8, colsample_bytree 0.8, up to 3000 trees with early stopping patience 100. LightGBM uses learning rate 0.03, num_leaves 63, feature_fraction 0.8, bagging_fraction 0.8, min_data_in_leaf 50, up to 5000 trees with early stopping patience 100. Early-stopping validation is taken as the last 10% of each training fold in time order. All training is time ordered with no shuffling and a fixed random seed for reproducibility.

### 3.2.2. Evaluation

This study uses rolling-origin splits with K = 8 folds. In each fold, the model is trained on all data up to a cutoff point and tested on the next contiguous window of approximately one calendar quarter. To assess cross-site generalization, a leave-one-site-out procedure is also applied, in which models are trained on three sites and tested on the held-out site using the same calendar spans. Performance is evaluated using MAE, RMSE, and R-squared. Pooled results are reported “micro” metrics, computed across all test samples from all sites.

### 3.2.3. Result

On the pooled data, boosted trees outperform linear baselines and Random Forests. As shown in Table 1, LightGBM achieves the best overall accuracy (RMSE 0.1611, MAE 0.1276, R-squared 0.4501), followed by XGBoost (RMSE 0.1678, R-squared 0.3533) and Random Forest (RMSE 0.1684, R-squared 0.3541). Linear models trail behind (for example, Ridge RMSE 0.1760, R-

DOI: 10.54254/2755-2721/2025.LD28453

squared 0.3350). Although the ranking is clear, pooled R-squared remains modest (about 0.335–0.450), indicating that aggregation underfits site-specific heterogeneity and motivating the per-site analysis in Section 3.3.

Table 1. Pooled performance (overall, micro averages)

| Model | MAE | RMSE | R-squared |
|---|---|---|---|
| Lasso | 0.1344 | 0.1747 | 0.3367 |
| LightGBM | 0.1276 | 0.1611 | 0.4501 |
| Linear | 0.1307 | 0.1725 | 0.3425 |
| Random Forest | 0.1277 | 0.1684 | 0.3541 |
| Ridge | 0.1361 | 0.1760 | 0.3350 |
| XGBoost | 0.1276 | 0.1678 | 0.3533 |

### 3.3. Per-site modeling

To expose heterogeneity that pooled training blurs, this study repeats the same training procedure independently for each site using the identical chronological splits and features_top27. Preprocessing, scaling, early-stopping setup, and fixed random seed mirror the pooled runs so that differences reflect site characteristics rather than configuration drift.

Site-specific modeling reveals two high-skill sites (Loc1 and Loc4) and two low-skill sites (Loc2 and Loc3) (Table 2). The best model for Loc1 is Random Forest, which yields low error and strong explanatory power. Loc4 is best modeled by XGBoost, again with high R-squared and the lowest errors across all sites. Loc2 and Loc3 remain challenging; LightGBM is best on both but R-squared is low, indicating that additional structure is needed beyond pooled features. These outcomes define the two optimization routes used later: seasonal expert models for high-skill sites and stepwise, physics-guided feature additions for low-skill sites.

Table 2. Per-site best models (hourly power as target)

| Site | Best Model | MAE | RMSE | R-squared |
|---|---|---|---|---|
| Loc1 | Random Forest | 0.1130 | 0.1471 | 0.7392 |
| Loc2 | LightGBM | 0.1679 | 0.2050 | 0.1462 |
| Loc3 | LightGBM | 0.1691 | 0.2038 | 0.1620 |
| Loc4 | XGBoost | 0.0755 | 0.1040 | 0.7934 |

Interpretation. Loc1 and Loc4 achieve R-squared around 0.74–0.79 with low MAE and RMSE, which supports treating them as a high-R-squared group and applying seasonal specialists (DJF, MAM, JJA, SON). Loc2 and Loc3 achieve R-squared around 0.15–0.16 even with the best model, pointing to cross-season fit limitations; we therefore apply a stepwise feature program that adds wind-speed squared terms, u/v vector winds, temperature-humidity interaction, and short lags, followed by a final pass that introduces hour-of-day and month-of-year to capture diurnal and seasonal cycles.

DOI: 10.54254/2755-2721/2025.LD28453

### 3.4. Feature importance and feature slimming

Procedure. Feature importance was computed using gradient-boosted trees under a leave-one-site-out (LOSO) scheme. For each LOSO fold, LightGBM and XGBoost were trained on the three-site training set using the full engineered feature set. Gain-based importances were extracted from both models, normalized to sum to one per fold, and then averaged across folds and across the two algorithms. To avoid overfitting to a single split or model, stability was required: a feature had to appear with nonzero gain in most folds before being ranked highly. A compact Top-27 list was then fixed by selecting the highest-ranked and most stable features, and this list was used for all subsequent experiments. All other training settings, data splits, and hyperparameters were kept the same as in the per-site runs so that performance differences could be attributed to the feature set rather than configuration drift.

Results with Top-27. Retraining each site with its previously selected best model shows that the Top-27 set preserves or slightly improves explanatory power while reducing model complexity. Three of the four sites see higher R-squared values, and error changes remain small. The per-site results with the Top-27 feature set are shown in Table 3.

Table 3. Per-site best models using the Top-27 feature set

| Site | Best Model | MAE | RMSE | R-squared |
|---|---|---|---|---|
| Loc1 | Random Forest | 0.1210 | 0.1394 | 0.7874 |
| Loc2 | LightGBM | 0.1679 | 0.2050 | 0.1762 |
| Loc3 | LightGBM | 0.1691 | 0.2038 | 0.1920 |
| Loc4 | XGBoost | 0.0951 | 0.1120 | 0.8056 |

As shown in Table 4, relative to full-feature configuration, the streamlined feature set yields modest performance shifts. R-squared increases by about +0.048 at Loc1, +0.030 at Loc2, +0.030 at Loc3, and +0.012 at Loc4. RMSE improves at Loc1 (−0.008) and is essentially unchanged at Loc2–Loc3; Loc4 shows a small increase (+0.008). MAE moves slightly at Loc1 (+0.008) and Loc4 (+0.020) and is unchanged at Loc2–Loc3. These trade-offs are consistent with the objective of slimming the feature space while keeping overall accuracy within noise levels.

Table 4. Top-27 minus full-feature deltas (Top-27 − full)

| Site | ΔMAE | ΔRMSE | ΔR-squared |
|---|---|---|---|
| Loc1 | +0.008 | −0.008 | +0.048 |
| Loc2 | +0.000 | +0.000 | +0.030 |
| Loc3 | +0.000 | +0.000 | +0.030 |
| Loc4 | +0.020 | +0.008 | +0.012 |

Interpretation. The Top-27 set matches or improves R-squared at all sites while reducing the number of input variables to a compact, stable subset. Loc1 benefits the most, with better variance explanation and lower RMSE. Loc2 and Loc3 gain modestly in R-squared without changing error magnitudes, which is desirable given their lower baseline skill. Loc4 exhibits a small trade-off between RMSE and R-squared, a pattern attributable to fold-to-fold variance and the sensitivity of RMSE to tail errors. Overall, the Top-27 shortlist retains predictive power and simplifies training

DOI: 10.54254/2755-2721/2025.LD28453

and deployment, so it is adopted for all subsequent seasonal-expert and stepwise feature experiments.

### 3.5. Cross-site generalization with LOSO (Top-27 features)

Design. To assess transferability across locations, the study evaluated the same best model class per site under LOSO using the Top-27 features. For each target site, models were trained on the other three sites and tested on the held-out site over the same calendar spans as in the per-site runs.

Results. LOSO generalization remains strong for the high-skill sites and improves materially for the low-skill sites (Table 5). Relative to the per-site Top-27 runs, R-squared increases by 0.056 at Loc1, 0.073 at Loc2, 0.089 at Loc3, and 0.055 at Loc4. MAE decreases by 0.015 at Loc1, 0.007 at Loc2, 0.014 at Loc3, and 0.018 at Loc4; RMSE decreases at three sites and is effectively unchanged at Loc2. These gains indicate that training on three sites provides useful diversity that stabilizes the learned mapping from meteorology to power, while the compact Top-27 set transfers well.

Table 5. LOSO performance by site (train on other three sites; test on the held-out site)

| Site | Best Model | MAE | RMSE | R-squared |
|---|---|---|---|---|
| Loc1 | Random Forest | 0.106 | 0.138 | 0.843 |
| Loc2 | LightGBM | 0.161 | 0.206 | 0.249 |
| Loc3 | LightGBM | 0.155 | 0.196 | 0.281 |
| Loc4 | XGBoost | 0.077 | 0.103 | 0.861 |

Interpretation. The Top-27 shortlist yields a compact, stable representation that supports both within-site modeling and cross-site transfer. High-skill sites retain high explanatory power under LOSO (R-squared about 0.84–0.86), while low-skill sites benefit most from pooled training across locations. These results justify adopting the Top-27 features for the subsequent seasonal-expert experiments and for the stepwise, physics-guided feature program.

### 3.6. High-R-squared group: seasonal expert models (Loc1, Loc4)

Setup. Using the Top-27 features and the per-site best model class (Loc1: Random Forest; Loc4: XGBoost), each site was split into DJF, MAM, JJA, and SON. Within each season the data remained in chronological order, early stopping used the last 10 percent of the training fold, and no information crossed seasonal boundaries. Evaluation compares each seasonal expert with the all-year baseline trained on the same Top-27 inputs.

Results (Table 6 and Table 7). Seasonal experts produce R-squared values that are close to the all-year model for both sites, with differences generally within a few hundredths. Winter and autumn are marginally higher than the annual model, while summer is slightly lower. Overall, the seasonal split does not materially change accuracy.

DOI: 10.54254/2755-2721/2025.LD28453

Table 6. Loc1 seasonal experts versus all-year baseline (Top-27 features)

| Season | R-squared | ΔR-squared vs All-year (0.843) |
|---|---|---|
| DJF | 0.852 | +0.009 |
| MAM | 0.844 | +0.001 |
| JJA | 0.816 | −0.027 |
| SON | 0.861 | +0.018 |
| All-year | 0.843 | — |

Table 7. Loc4 seasonal experts versus all-year baseline (Top-27 features)

| Season | R-squared | ΔR-squared vs All-year (0.843) |
|---|---|---|
| DJF | 0.869 | +0.008 |
| MAM | 0.865 | +0.004 |
| JJA | 0.852 | −0.009 |
| SON | 0.875 | +0.014 |
| All-year | 0.861 | — |

Interpretation. The seasonal experts deliver only modest changes relative to the all-year model. This limited impact is expected because the meteorological predictors themselves are seasonally structured (for example, wind regimes, temperature, and humidity), allowing the models to absorb much of the seasonal signal without explicit seasonalization. In practice, the annual model is adequate for these high–R-squared sites, with seasonal experts offering small gains in winter and autumn but no consistent advantage overall.

## 3.7. Low-R-squared group: stepwise features and seasonal validation (Loc2, Loc3)

### 3.7.1. Stepwise feature program (A–D)

For Loc2 and Loc3, a controlled, add-one-thing-at-a-time feature program was applied on top of the Top-27 inputs using the same LightGBM configuration and time-aware splits. Step-A adds wind-speed squared terms to capture the nonlinearity of the partial-load regime. Step-B replaces or complements speed–direction with u/v vector winds to respect circular geometry. Step-C introduces a thermo-humidity interaction as a proxy for density effects. Step-D adds short lags of Power, wind speed, temperature and humidity at 1, 3 and 6 hours; lags are created after the split and within site only to avoid leakage. Each step is evaluated against the same baseline using identical folds; we record ΔR-squared, ΔMAE and ΔRMSE and keep a cumulative “gain” curve. In aggregate, the most reliable improvement comes from Step-D, with Steps-A and B providing small but consistent lifts and Step-C giving mixed effects. The corresponding validation performance of the LightGBM model at Loc2 and Loc3 is summarized in Table 8. These patterns suggest that short-term persistence and correct wind representation matter more than additional thermo-humidity structure for these two sites. Because the gains are incremental and site-specific, only the steps that improve validation metrics consistently across folds are retained.

DOI: 10.54254/2755-2721/2025.LD28453

Table 8. Validation results of LightGBM at Loc2 and Loc3

| Site_ID | Model | MAE | RMSE | R2 |
|---|---|---|---|---|
| Loc2 | LightGBM | 0.084 | 0.117 | 0.684 |
| Loc3 | LightGBM | 0.097 | 0.098 | 0.713 |

### 3.7.2. Seasonal validation

The study next tested whether splitting the low-skill sites by season provides additional benefit. Using the same model class (LightGBM) and Top-27 inputs, one model was trained per season (DJF, MAM, JJA, SON) and compared each to the all-year baseline. Seasonalization yields only limited changes in accuracy, with modest gains in winter and autumn and small degradations in summer. The detailed validation results for Loc2 and Loc3 are summarized in Table 9 and Table 10. This indicates that the meteorological predictors already carry much of the seasonal signal, so explicit seasonal splits do not materially enhance generalization for these sites.

Table 9. Loc2 seasonal experts versus all-year baseline (Top-27 features; LightGBM; annual MAE 0.084, RMSE 0.117)

| Season | R-squared | ΔR-squared vs All-year (0.684) |
|---|---|---|
| DJF | 0.706 | +0.022 |
| MAM | 0.698 | +0.014 |
| JJA | 0.673 | −0.011 |
| SON | 0.725 | +0.041 |
| All-year | 0.684 | — |

Table 10. Loc3 seasonal experts versus all-year baseline (Top-27 features; LightGBM; annual MAE 0.097, RMSE 0.098)

| Season | R-squared | ΔR-squared vs All-year (0.684) |
|---|---|---|
| DJF | 0.718 | +0.005 |
| MAM | 0.725 | +0.012 |
| JJA | 0.692 | −0.021 |
| SON | 0.706 | −0.007 |
| All-year | 0.713 | — |

Seasonal splits change performance by only a few hundredths and do not alter the overall conclusion for the low-R-squared group: cross-season generalization is the limiting factor, and the most effective remedy is the stepwise, physics-guided features—especially short lags and vector winds—rather than seasonalization. Consequently, for Loc2 and Loc3, the study proceeds with the subset of steps that provide consistent ΔR-squared and error reductions, followed by a final pass that adds hour-of-day and month-of-year to capture residual diurnal and seasonal structure.

DOI: 10.54254/2755-2721/2025.LD28453

### 3.8. Time features applied uniformly across all sites

After completing site-specific optimization (seasonal experts for Loc1 and Loc4; stepwise features for Loc2 and Loc3), two calendar variables were appended to each site's feature set: hour_of_day (0–23) and month_of_year (1–12). To prevent leakage, these variables were derived from timestamps after the train–test split within each fold. For tree models one-hot encodings were used (24 columns for hours, 12 for months), while for linear models, cyclical encodings with sine and cosine pairs for hour and month were employed. No other changes were made to the data pipeline, splits, or hyperparameters. Performance was evaluated with the same time-aware protocol as before. The baseline for comparison is the best configuration per site under LOSO with the Top-27 features (Sections F–G).

Adding the two time variables yields consistent and substantial accuracy gains at all sites. The detailed results are summarized in Tables 11 and 12. Errors decrease sharply and R-squared rises into the 0.98 range, indicating that explicit diurnal and seasonal signals complement meteorological predictors and help the models capture residual structure.

Table 11. Final performance with hour_of_day and month_of_year added (per-site best model)

| Site | RMSE | MAE | R-squared |
|---|---|---|---|
| Loc1 | 0.03223 | 0.02340 | 0.98770 |
| Loc2 | 0.02662 | 0.01883 | 0.98001 |
| Loc3 | 0.03275 | 0.02378 | 0.97813 |
| Loc4 | 0.02517 | 0.01714 | 0.98842 |

Table 12. Improvement relative to the LOSO Top-27 baseline (Δ = final − baseline)

| Site | ΔRMSE | ΔMAE | ΔR-squared |
|---|---|---|---|
| Loc1 | −0.10577 | −0.08260 | +0.14470 |
| Loc2 | −0.17938 | −0.14217 | +0.73101 |
| Loc3 | −0.16325 | −0.13122 | +0.69713 |
| Loc4 | −0.07783 | −0.05986 | +0.12742 |

The hour and month variables provide a compact way to encode persistent daily and seasonal cycles that are not fully captured by the meteorological inputs alone. The gains are uniform across all four sites, with the largest improvements at the two previously low-R-squared sites. Given the minimal modeling overhead and strong benefits, these time features are retained in the final specification for all sites.

## 4. Results

### 4.1. Pooled model comparison

On the pooled dataset, gradient-boosted trees outperform linear baselines and Random Forest. LightGBM achieves the best overall accuracy (RMSE 0.1611, MAE 0.1276, R-squared 0.4501), followed by XGBoost (RMSE 0.1678, R-squared 0.3533) and Random Forest (RMSE 0.1684, R-squared 0.3541). Linear models trail behind (for example, Ridge RMSE 0.1760, R-squared 0.3350).

DOI: 10.54254/2755-2721/2025.LD28453

Although the ranking is clear, pooled R-squared remains modest (about 0.335–0.450), indicating that aggregation underfits inter-site differences (Table 1).

### 4.2. Per-site performance

Training independently by location exposes pronounced heterogeneity. Loc1 and Loc4 are high-skill sites with low errors and R-squared of 0.7392 and 0.7934, respectively. Loc2 and Loc3 are low-skill sites; LightGBM performs best but yields R-squared of only 0.1462 and 0.1620 (Table 2). These results motivate site-aware optimization.

### 4.3. Feature importance and slimming

Gain-based importances from LightGBM/XGBoost under leave-one-site-out produce a stable Top-27 shortlist. Retraining each site's best model with Top-27 preserves or improves explanatory power while reducing dimensionality. Relative to full features, R-squared increases by +0.048 (Loc1), +0.030 (Loc2), +0.030 (Loc3), and +0.012 (Loc4) with negligible changes in MAE/RMSE. The compact feature set is retained for subsequent experiments.

### 4.4. Cross-site generalization (LOSO)

With Top-27 inputs and the site-specific best model class, leave-one-site-out improves generalization. R-squared increases to 0.843 (Loc1), 0.249 (Loc2), 0.281 (Loc3), and 0.861 (Loc4); MAE also decreases (Table 5). Training on three sites supplies diversity that stabilizes the mapping from meteorology to power at the held-out site.

### 4.5. High-R-squared group: seasonal experts

For Loc1 and Loc4, seasonal experts (DJF, MAM, JJA, SON) perform similarly to the all-year model, with small gains in winter and autumn and slight dips in summer. Loc1 ranges from 0.816 to 0.861 around an annual value of 0.843; Loc4 ranges from 0.852 to 0.875 around 0.861 (Tables 6–7). Overall impact is limited, consistent with seasonal information already embedded in the meteorological predictors.

### 4.6. Low-R-squared group: stepwise features and seasonal check

On Loc2 and Loc3, the stepwise feature program shows that short lags and vector winds provide the most consistent gains, wind-speed squared offers small improvements, and the thermo-humidity interaction has mixed effects. Seasonalization changes accuracy by only a few hundredths (modest gains in DJF/SON; small degradations in JJA), confirming that seasonal splits alone do not resolve cross-season generalization (Tables 9–10).

### 4.7. Time features added uniformly

After appending hour_of_day and month_of_year to all sites, accuracy improves sharply and uniformly. Final R-squared reaches 0.9877 (Loc1), 0.9800 (Loc2), 0.9781 (Loc3), and 0.9884 (Loc4), with RMSE between 0.025 and 0.033 (Table 11). Relative to the LOSO Top-27 baselines, ΔR-squared is +0.1447 (Loc1), +0.7310 (Loc2), +0.6971 (Loc3), and +0.1274 (Loc4) (Table 12).

DOI: 10.54254/2755-2721/2025.LD28453

Explicit time encodings therefore capture residual diurnal and seasonal structure not fully learned from meteorological inputs alone.

## 5. Discussion

### 5.1. Principal findings

Three conclusions stand out. First, site heterogeneity is material: pooled training underfits inter-site differences, while per-site modeling and LOSO clarify performance and transferability. Second, a compact and stable feature core is sufficient: the Top-27 shortlist preserves or improves skill while simplifying the pipeline. Third, lightweight time encodings are highly effective: adding hour and month delivers the largest gains, especially at the previously low-skill sites.

### 5.2. Interpreting seasonal effects

Seasonal experts yield only marginal changes at high-skill sites because meteorological variables already encode seasonal regimes. Where improvements appear (primarily in winter and autumn), they are small and do not justify the added operational complexity. For low-skill sites, seasonalization is not the bottleneck; cross-season generalization and short-horizon dynamics dominate.

### 5.3. Factors that contribute to low skilled sites

Persistence and wind representation matter most. Short lags exploit local temporal memory that raw meteorology does not fully capture, and u/v vectors avoid angular discontinuities inherent in directional encoding. These choices provide steady, interpretable gains without heavier architectures.

### 5.4. Operational implications

A practical specification emerges. For high-R-squared sites, deploy the per-site best tree model with the Top-27 features plus hour/month; seasonal experts are optional. For low-R-squared sites, retain the Top-27 and add short lags and vector winds, again including hour/month as standard. Leave-one-site-out should be part of evaluation whenever models are intended to transfer across locations.

### 5.5. Limitations

Findings are based on four sites with hourly resolution and normalized power; performance may differ with other assets or sampling rates. In deployment, inputs will come from weather forecasts rather than observations, which typically reduces skill. The robust power-curve filter may remove rare but valid extremes, and short-gap interpolation can smooth variability. Although leakage safeguards were used (deriving time features and lags after splitting), the strong improvements from time variables should be replicated on additional sites and years.

### 5.6. Future work

Two immediate extensions are planned: calibrated probabilistic forecasting to quantify uncertainty and regime-aware or mixture-of-experts models keyed to stability or synoptic classes. Further work with forecast-driven inputs (e.g., NWP ensembles), multi-horizon objectives, and larger site networks via transfer learning may enhance generalization while maintaining a compact feature set.

DOI: 10.54254/2755-2721/2025.LD28453

## 6. Conclusion

This study offers a deployment-minded blueprint for short-term wind power forecasting across heterogeneous sites. Through pooled vs. per-site experiments and LOSO evaluation, the analysis demonstrates that a compact, physics-aware feature core (Top-27) augmented with lightweight calendar encodings—and, for low-skill sites, short persistence lags and u/v wind vectors—yields consistent and interpretable gains without resorting to seasonal experts or heavier architectures. Operationally, per-site tree models with Top-27 plus hour/month suffice for high-skill assets; for low-skill sites, adding short lags and vector winds is recommended, and LOSO should remain standard when transfer across locations is intended. While the analysis relies on four sites, hourly resolution, and observed meteorology, the workflow is readily portable to forecast-driven inputs, larger networks, and multi-horizon objectives. Future extensions in probabilistic calibration, regime-aware mixtures, and transfer learning with NWP ensembles can further improve reliability and scalability. By emphasizing lean features, rigorous leakage-safe evaluation, and clarity of operational choices, the paper advances a practical path to trustworthy wind-asset forecasting.

## References

[1] Olson, J. B., Kenyon, J. S., Angevine, W. M., Brown, J. M., Pagowski, M., & Suselj, K. (2019). Improving wind energy forecasting through numerical weather prediction model development. Bulletin of the American Meteorological Society, 100(11), 2201–2220. https: //doi.org/10.1175/BAMS-D-18-0040.1

[2] Benjamin, S. G., Weygandt, S. S., Brown, J. M., Hu, M., Alexander, C. R., Smirnova, T. G., Olson, J. B., James, E. P., Dowell, D. C., Grell, G. A., Lin, H., Manikin, G. S., Sponseller, D. W., Smith, T. L., Moninger, W. R., Kenyon, J., & Jamison, B. D. (2016). A North American hourly assimilation and model forecast cycle: The Rapid Refresh. Monthly Weather Review, 144(4), 1669–1693. https: //doi.org/10.1175/MWR-D-15-0242.1

[3] Skamarock, W. C., Klemp, J. B., Dudhia, J., Gill, D. O., Liu, Z., Berner, J., Wang, W., Powers, J. G., Duda, M. G., Barker, D. M., & Huang, X.-Y. (2019). A description of the Advanced Research WRF model version 4 (NCAR Technical Note NCAR/TN-556+STR). National Center for Atmospheric Research. https: //opensky.ucar.edu/islandora/object/technotes%3A500

[4] Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., … Thépaut, J.-N. (2020). The ERA5 global reanalysis. Quarterly Journal of the Royal Meteorological Society, 146(730), 1999–2049. https: //doi.org/10.1002/qj.3803

[5] Draxl, C., Hodge, B.-M., Clifton, A., & McCaa, J. (2015). Overview and meteorological validation of the Wind Integration National Dataset (WIND) Toolkit (NREL/TP-5000-61740). National Renewable Energy Laboratory. https: //www.nrel.gov/docs/fy15osti/61740.pdf

[6] Foley, A. M., Leahy, P. G., Marvuglia, A., & McKeogh, E. J. (2012). Current methods and advances in forecasting of wind power generation. Renewable Energy, 37(1), 1–8. https: //doi.org/10.1016/j.renene.2011.05.033

[7] Meng, W., Li, C., Wang, J., Zhang, Y., & Zhao, C. (2022). Short-term wind power forecasting based on mixed data sampling and feature engineering. Energy, 244, 123195. https: //doi.org/10.1016/j.energy.2021.123195

[8] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). https: //doi.org/10.1145/2939672.2939785

[9] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In Advances in Neural Information Processing Systems 30 (NeurIPS 2017) (pp. 3146–3154). http: //papers.nips.cc/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html

[10] Perr-Sauer, J., Optis, M., Fields, M. J., Philips, C., Debnath, U., Martin, W., Letchford, J., & King, R. (2021). OpenOA: An open-source codebase for operational analysis of wind farms. Journal of Open Source Software, 6(58), 2171. https: //doi.org/10.21105/joss.02171

[11] Morrison, R., Polikarpova, I., Infield, D., & Carroll, J. (2022). Anomaly detection in wind turbine SCADA data for power curve cleaning. Renewable Energy, 191, 701–717. https: //doi.org/10.1016/j.renene.2022.04.114

[12] Yao, Q., Hu, Y., Wang, J., Yang, C., & Li, Y. (2023). A composed method of cleaning anomaly data for wind turbine SCADA. Renewable Energy, 211, 111–123. https: //doi.org/10.1016/j.renene.2023.04.044

DOI: 10.54254/2755-2721/2025.LD28453

[13] Mardia, K. V., & Jupp, P. E. (2000). Directional statistics. John Wiley & Sons. https://onlinelibrary.wiley.com/doi/book/10.1002/9780470316979

[14] Wang, Z., Wang, Y., Ren, H., Zhang, L., Zhang, H., & Liang, J. (2024). Spatiotemporal wind speed forecasting using u-v components and graph attention networks. Scientific Reports, 14, 39519. https://www.nature.com/articles/s41598-024-39519-9

[15] Picard, A., Davis, R. S., Gläser, M., & Fujii, K. (2008). Revised formula for the density of moist air (CIPM-2007). Metrologia, 45(2), 149–155. https://doi.org/10.1088/0026-1394/45/2/004

[16] Hyndman, R. J., & Athanasopoulos, G. (2021). Forecasting: Principles and practice (3rd ed.). OTexts. https://otexts.com/fpp3/

[17] Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: An analysis and review. International Journal of Forecasting, 16(4), 437–450. https://doi.org/10.1016/S0169-2070(00)00065-0

[18] Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Hartig, F., & Dormann, C. F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography, 40(8), 913–929. https://doi.org/10.1111/ecog.02881

[19] Hong, T., Pinson, P., Fan, S., Zareipour, H., Troccoli, A., & Hyndman, R. J. (2016). Probabilistic energy forecasting: Global Energy Forecasting Competition 2014 and beyond. International Journal of Forecasting, 32(3), 896–913. https://doi.org/10.1016/j.ijforecast.2016.02.001

[20] Yang, X., Vano, J. A., Feldstein, S. B., Wu, Y., & Tippett, M. K. (2024). Skillful seasonal prediction of wind energy resources in the U.S. Great Plains. Communications Earth & Environment, 5, 263. https://doi.org/10.1038/s43247-024-01457-w

[21] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems 30 (NeurIPS 2017) (pp. 4765–4774). https://papers.nips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf